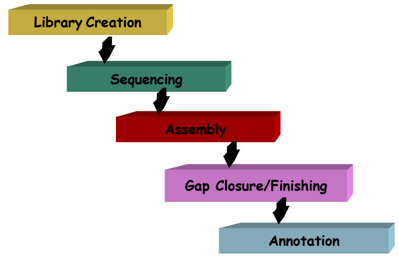
High Quality Microbial Finishing at JGI

Alla Lapidus (alapidus@lbl.gov)[1], Eugene Goltsman[1], Steve Lowry[1], Hui Sun[1], Alicia Clum[1], Stephan Trong[1], Pat Kale[1], Alex Copeland[1], Patrick Chain[2], Cliff Han[3], Tom Brettin[3], Jeremy Schmutz[4], Paul Richardson[1]

[1] DOE JGI-PGF, Walnut Creek, CA; [2] JGI-LLNL, Livermore, CA; [3] JGI-LANL, Los Alamos, NM; [4] JGI-Stanford, Stanford, CA

The value of complete microbial genome sequence is established and appreciated by scientific community. A finished genome represents the genome assembly of high accuracy and quality (with no gaps), verified and confirmed through a number of computer and lab experiments. Several yeas ago JGI has established a set of high standards for the final microbial assembly and has been strictly following them thereafter. More than 100 microbial projects have been completed since that time within the framework of the JGI's portfolio (DOE GTL program, DOE Microbial program and the Community Sequencing Program). Progress in DNA sequencing technology, design of new vectors for library construction, improvements in finishing strategies and tools, as well as the availability of a number of assemblers and advanced methods for OFR finding and genome annotation have significantly reduced the time required for genome closure. Despite this fact, complexity and speed of genome closure depends on the quality of DNA received, the whole genome shotgun libraries produced from this DNA, GC content of the genome, the size and frequency of identical or nearly identical repetitive structures, and the amount of regions that can not be cloned or hard to clone in E.coli. The whole genome finishing/assembly improvement pipeline will be presented showing the lab approaches and computational finishing techniques developed and implemented at JGI for finishing the large number of microbial projects in the queue. We also will present our progress in completing metagenomic projects. A number of projects for which the combination of different sequencing technologies (Sanger and 454) and finishing strategies were used will also be presented.


Library Creation

Sequencing

Assembly

Gap Closure/Finishing

Annotation

Finished GTL projects

- *Desulforudis audaxviator/ (Desulfotomaculum-like organism)*
- *Pelobacter carbinolicus DSM 2380*
- *Pelobacter propionicus DSM 2379*
- *Rhodoferax ferrireducens DSM 15236*
- *Desulfovibrio vulgaris DePue*
- *Korarchaeota Community*

GTL projects in progress

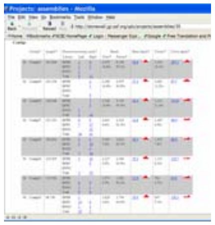
- *Geobacter uraniumreducens Rf-4*
- *Geobacter bemidjiensis Bem(T)*

Finishing Steps

- Automated repeat resolution
- 1st round of primer walk (short gaps closed)
- 2d round of primer walk (gaps ~ 4kb closed)
- Manual repeat resolution while closing other captured gaps (4kb < L < 6kb)
- PCRs to map contigs and cover ucaptured gaps
- Quality improvement (1x regions; poor quality; 454 resequencing)
- Assembly QC

Project analysis prior finishing

•http://stonewall.jgi-psf.org/qdx/projects/



Quality: JGI Standards for finished microbes

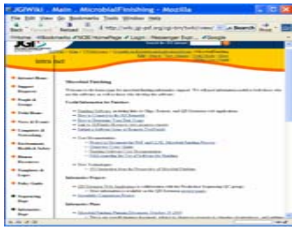
- All low-quality areas (<Q30) are reviewed and re-sequenced
- No single-read coverage is permitted in Sanger only areas
- "454 contigs" should be assigned quality Q20 until real quality assignment becomes available. Newbler's quality scores that fall below Q20 should be preserved; overlapping parts of "454 contigs" should be assigned quality 7
- Locate all areas covered with 454 only zero Sanger coverage). PCR each 454 only area and sequence resulting PCR products by standard primer walking (Sanger)
- Up to 5%* of a genome can remain covered with pyrosequence only
- All remaining 454-only regions should be tagged: "No confirming data from Sanger sequence was collected, 454 only data"
- Locate "454+sanger" regions where consensus quality falls below Q30, and/or there is a discrepancy between the two technologies. At least one attempt should be made to cover such areas with at least one high quality read (Q>20). Unresolved problems should be tagged
- All positions where an aligned high-quality read (>Q29) disagrees with the consensus base should be visually inspected. If unable to resolve, additional chemistries should be attempted to resolve discrepancies or tag the area
- All strings of xxxx are resolved in the final sequence.
- All repeats are verified
- The ends of final contigs (chromosomes, plasmids) are checked
- The final error rate in consensus must be not higher than 1 per 50 Kb.
- The final assembly is given a QA/QC check

•(*) –this percentage depends on the 454 software improvement and can grow with time

Microbial assembly QC procedure

- Import shotgun traces, finishing traces, subassemblies and finished sequence from JGI servers.
- Build an assembly with all shotgun reads for reference.
- Place all shotgun read pairs on finished consensus and build QC assembly with only good placed pairs. Check clone junctions and sort large repeats.
- Pull in finishing traces and leftover shotgun traces for every remaining low quality or single subclone area and rebuild the assembly.
- Computationally finish the QC assembly and export a new consensus.
- Compare the QC consensus against the original finished consensus, for any discrepancies examine the position in the finished ace file and the QC assembly.
- Validate each repeated segment in consensus against valid placed shotgun read pairs.

SOP

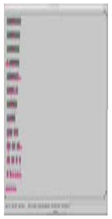


New technologies in finishing

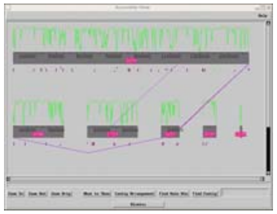
Great tool for contigs scaffolding. Especially useful in case of low GC genomes
High GC% (>65%) – sequencing problems
Low GC% (<40%) – problems in cloning (biased)

Assembly Views of Halothermothrix orenii

Repeat resolution + primer walks



(Repeat resolution + primer walks) + **454 run**



Closed Microbial projects

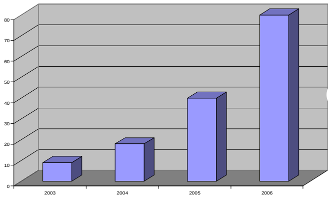


Fig1. Amount of microbial projects finished by JGI per year

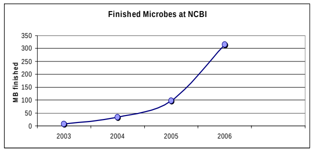


Fig2. Amount of finished sequence (in Mb) submitted by JGI to NCBI per year

Development

- **Cloning strategy improvement**
- **Optimized Assembler selection**
- **Software development and improvement for:**

~ process automation
~ repeat identification and characterization
~ visualization of the process
~ genome comparison based finishing
~ annotation based genome finishing
~ lab experiments prediction and planning
~ more robust polishing (automation, new technologies)

- **Diversity of lab approaches for difficult template amplification and sequencing**
- **New sequencing technologies implementation (454, Solexa)**

Environmental genomics:

- extract DNA from an environmental sample
- sequence and analyze

- **DNA from fracture water collected at 2.8 km depth was sequenced and assembled into a single complete chromosome. This new Gram-positive genus was named *Candidatus Desulforudis audaxviator*. The genome content of *Desulforudis audaxviator* supports the geochemical analysis that indicates sulfate instead of oxygen is used for the generation of energy.**

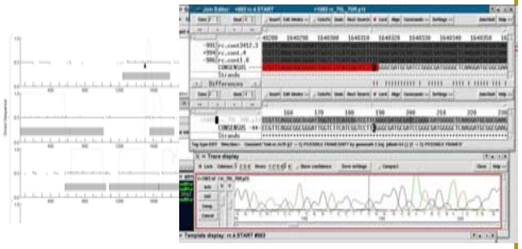
- ***Korarchaeota* Community:**

Contains an organism that represents what could be one of the least evolved lineages of modern life that has been detected in nature so far. The enrichment community has been phylogenetically characterized and is known to comprise a relatively low diversity of other hyperthermophilic archaea and bacteria.

Annotation related tools in finishing:

GeneMark

Frame shift correction



•This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396. LBNL - 62254 Poster